Assessment and Prediction of Organic Pollution Index Using Deep Learning Models: A Case Study of Major Lakes in Dhaka City, Bangladesh

Md Abdul Mutalib
Dept. of Computer science and Engineering
North East University Bangladesh
Sylhet-3100, Bangladesh
mutalib.cs@gmail.com

Md. Anamul Hossain Chowdhury
Dept. of Civil Engineering
Sylhet Engineering College
Sylhet-3100, Bangladesh
anamulchowdhury@sec.ac.bd

Rayhan Mahmud Rakib
Dept. of Civil Engineering
Sylhet Engineering College
Sylhet-3100, Bangladesh
rayhanrakib93415@gmail.com

Rumi Chowdhury
Dept. of Civil Engineering
Barishal Engineering College
Barishal, Bangladesh
rumichowdhury412@gmail.com

Saeid Islam Tanvir
Dept. of Civil Engineering
Sylhet Engineering College
Sylhet, Bangladesh
saeidislamtanvir@gmail.com

Md Abdur Rashid Sarker
Dept. of Ag. Construction and Environ. Eng
Sylhet Agricultural University
Sylhet, Bangladesh
rashidsarker.acee@sau.ac.bd

Md. Fahad Jubayer
Dept. of Food Engineering and Technology
Sylhet Agricultural University
Sylhet, Bangladesh
Jubayer.fet@sau.ac.bd

***Abstract*—Dhaka City has experienced rapid urbanization that has polluted its major lakes organically and urgent need of the hour are accurate assessment and prediction methods to handle water management sustainably. This current study evaluates and forecasts the Organic Pollution Index (OPI) by using deep learning techniques from the analysis of water quality data of three lakes (Dhanmondi, Hatirjheel and Gulshan) pH, DO, BOD, COD, TDS, Chloride, Turbidity, SS, and Total Alkalinity TDS, EC, collected over six years (2019-2024). Our models are Convolutional Neural Network (CNN), TabNet, a combination of CNN – LSTM network and XGBoost: CNN gives the best results (R^2= 0.9969, MAE= 0.1888). Three different models (weighted CNN, TabNet, and XGBoost) combination represented in the form of an ensemble model yields the best trade-offs between calibration ability (R2=0.9800) and robustness while the importance of features can be easily interpreted in the case of the TabNet. Results show the effectiveness of deep learning when used to capture nonlinear relationships to predict OPI using a scalable tool for real-time monitoring of pollution.**

***Index Terms*—Organic Pollution Index (OPI), deep learning, CNN, CNN-LSTM, TabNet, XGBoost, water quality prediction, Dhaka City lakes**

I. Introduction

The rapid unmanaged urban expansion combined with demographic increases in Dhaka City has caused major damage to all natural water resources including lakes ponds and wetlands. Natural water bodies experience rising dangers because of pollution along with encroachment and enrichment from untreated industrial and domestic wastewater [1]. Organic pollution measurements are determined by the OPI since such contamination affects aquatic environments while posing safety risks [2]. Water management sustains itself when organic pollution gets properly assessed. The OPI is a commonly employed tool for this purpose however classical assessment approaches cannot generate prompt forecasts nor meet current needs. The urban lakes of Dhaka that include Gulshan, Dhanmondi, and Hatirjheel serve essential functions in drainage management and temperature control and biodiversity support and public recreational activities. The lakes have become victims of intense ecological damage caused by organic contamination [3]. The OPI assessment of these water bodies enables scientists to track pollution intensity through numeric measurement systems and confirm environmental trends caused by urban pollution sources. Localized pollution control measures need this information to guide them for successful implementation. People now use artificial intelligence systems that apply machine learning (ML) techniques to predict water quality in terms of their capability to process the complex and voluminous datasets [4]. Due to its high dependency on training data, the model often demonstrates challenges in its robustness performance model [5]. The deep learning (DL) models produces better predictions at the cost of high complexity and extreme computational expenses [6]. DL models such as Neural Network and CNN outperform over traditional ML and hybrid models, especially in recognizing complex nonlinear relationships. This research constitutes the initial effort to evaluate and forecast organic pollution levels in Dhaka's lakes through the implementation of optimized deep

learning algorithms. The research applies CNN and DNN in a new way to OPI prediction while determining the major contributing factors to OPI. The issue of organic pollution in the lakes within Dhaka city remains unexplored in existing academic literature [7]. The study's novelty lies in its focus on Dhaka city lakes through deep learning prediction models which have never been used previously. The study delivers crucial information to environmental managers who can use it for developing data-based pollution control approaches specific to particular locations. The research helps achieve advanced sustainable water pollution management in cities through successful prediction of pollution drivers coupled with water quality trend analysis.

A. Objectives

- Assess organic pollution levels in Dhaka's major lakes using the OPI.
- Analyze key physicochemical factors influencing OPI through correlation analysis.
- Develop deep learning models (CNN, TabNet, LSTM, and Ensemble) to predict OPI from water quality data.
- Provide insights for sustainable water management based on model predictions.

II. Literature Review

The growing trend of untreated domestic and industrial waste discharge in urban water bodies represents a significant concern for Dhaka City as it develops rapidly [8]. Biodegradable organic compounds together with nitrates and phosphates appear in most wastewater discharges. Excessive pollution can cause harm to aquatic life even though natural water bodies perform limited purification of pollution. To successfully manage water quality people must identify what causes pollution alongside its consequences [9]. Researchers widely employ the OPI to measure organic load by combining BOD, COD, NH-N, and DO assessment parameters in freshwater systems [10]. The application of Machine learning models in artificial intelligence systems improved water quality prediction efficiency along with accuracy levels [11]. Traditional machine learning models display insufficient capability to identify intricate temporal along with nonlinear water quality patterns within data streams which leads to decreased predictive accuracy [12]. The deep learning models including CNN-LSTM have demonstrated better performance by delivering high accuracy in prediction for dynamic water systems which include South Korea's Nakdong River Basin [13]. Water quality predictability improves when CNN, LSTM and attention mechanism combinations operate as hybrid deep learning models according to Mi et al. [14]. These models extract complex temporal patterns that enable high NSE and low RMSE/MAE estimation results covering both rivers and lakes datasets. Forecasting water quality of Yangtze River by Cheng Zhang et al. using CNN and LSTM networks reached an RMSE and MAE as low as 1.08% and 6.8% [15]. Research conducted by Zeyang Wei et al. proved that DNN achieved high performance result of 93.37% and a Kappa coefficient of 0.902 for the lakes located in Wuhan, while for lakes in Huangshi, it reached 96.39% accuracy and a Kappa of 0.951 when used to forecast two lake water quality in China surpassing established predictive tools [16]. Studying organic pollution modeling through DL methods for South Asian urban lakes remains a major knowledge deficit in scientific research. Research on water quality prediction predominantly focuses on rivers and rural bodies instead of targeting the pollution found in urban lakes of Dhaka. The research fills this knowledge gap through optimized deep learning model applications for OPI prediction in Dhaka's lakes in order to enhance early warning systems and urban water management decisions.

III. Study Area

Dhaka is the capital city of Bangladesh, covering a total area of 306.38 square kilometers [17]. The urban metropolitan region contains different water reservoirs combined with lakes and open rivers. The study addresses three lakes located in Dhaka named Dhanmondi Lake and Hatirjheel Lake and Gulshan Lake. A map depicts the exact positions of these lakes in Fig. 1 The research used 864 water quality samples collected from the government website of Bangladesh which is Department of Environment, during its 6-year span from 2019 to 2024. Nine key physicochemical parameters were evaluated, encompassing pH, DO, BOD, COD, TDS, SS, total alkalinity, and electrical conductivity (EC). Samples were collected from various locations in the lakes chosen due to their significance and suitability for reliable OPI measurement.

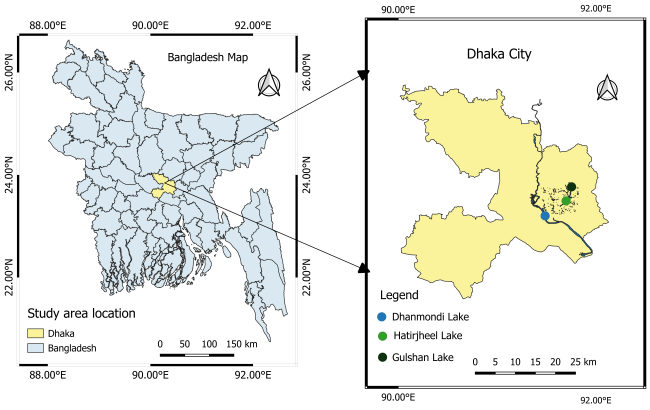


Fig. 1. Study area map.

IV. Methodology

The approach applied in this research incorporates a systematic and strict method towards OPI prediction using physicochemical parameters of water quality through machine learning and deep learning models. The experimental process is structured into six key stages: data preprocessing and feature engineering, CNN design, TabNet implementation, development of hybrid CNN-LSTM model, training of XGBoost Regressor, and soft ensemble strategy development. All of

these stages are designed and validated to offer the optimal prediction performance as demonstrated in Fig. 2.

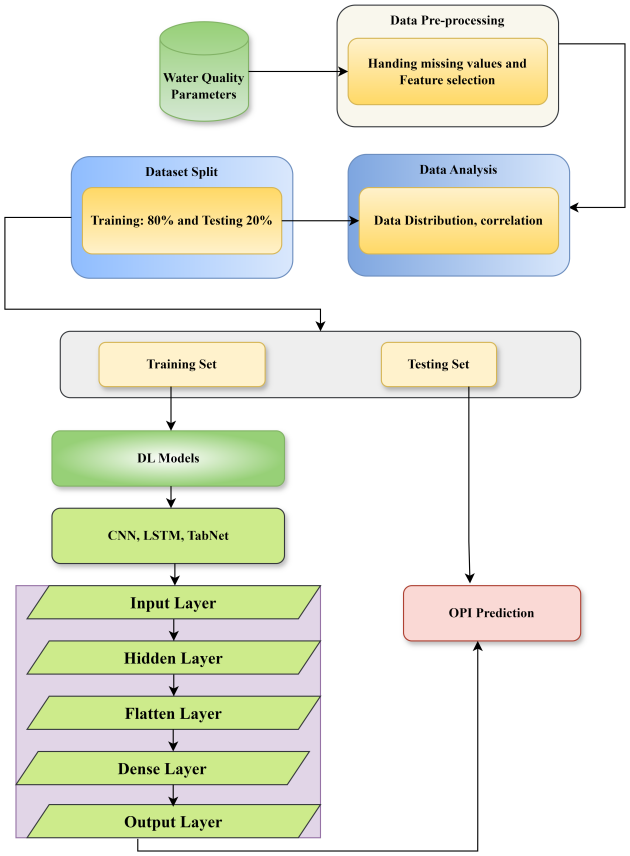


Fig. 2. Workflow diagram.

A. Data Cleaning and Feature Transformation

The data used in this research are collected from three lakes in Dhaka city and each sample has ten features such as the target variable OPI. The features include pH, DO, BOD, COD, TDS, chloride, suspended solids (SS), total alkalinity, and electrical conductivity (EC). Descriptive statistics of water quality parameters is presented in Table I. The data includes 864 entries with ten measurable variables. The dataset records average DO, BOD, and COD values – 5.6, 18.7 and 72.7 mg/L, respectively, which indicate significant organic pollution due to that neutral pH, that is, 7.4. TDS and EC present interchanged values, with mean concentrations of 221.5 mg/L and 426.4 µS/cm, respectively, with peak BOD (80 mg/L) and COD (343 mg/L) and SS (441.1 mg/L).The OPI values range from 0.7 to 84.8, which has an average of 17.2, indicating vary wide base of pollution levels.

Data pre-processing is a crucial process in any machine learning activity. In this study, categorical features that were impossible to transform into meaningful numeric values were excluded. 'Turbidity' and 'Location' features were dropped because they contained nominal values with high variance and low explicit correlation with OPI. For preserving temporal

TABLE I
DESCRIPTIVE STATISTICS OF WATER QUALITY PARAMETERS

	PH	DO	BOD	COD	TDS	Cl	SS	T.Alk	EC	OPI
count	864	864	864	864	864	864	864	864	864	864
mean	7.4	5.6	18.7	72.7	221.5	38	66.5	159.1	426.4	17.2
min	2.9	0.1	1.4	7	14	5.5	1	22	118.6	0.7
25%	7.3	3.6	8.6	37	146	28	19	112	311.8	8.5
50%	7.4	5.3	18	66	225.5	37	47.5	150	430	15.8
75%	7.6	7	25.4	100	276	47	92.3	202.3	528.3	24.1
max	9.4	50	80	343	487	238	441.1	352	855	84.8
std	0.3	3.1	11.8	42.3	83.2	15.6	60.8	60.5	129.1	10.5

information, the 'Date' feature was transformed into three numerical features: year, month, and day, keeping seasonal and periodic patterns without unnecessarily increasing dimensionality. Prior to model training, the data were preprocessed by standardizing all numeric features via the z-score method. The scikit-learn's StandardScaler was utilized for this to provide each feature with zero mean and one standard deviation. Standardization was crucial in supporting convergence rates of training and is crucial as it ensures that gradient updates are normalized in a consistent manner, particularly with deep learning architectures where unnormalized data may cause unstable weights updates [18].

The histograms Fig. 5 indicate that most water quality parameters are distributed to right skew, with BOD being 10–20mg/L and up to 80mg/L, while COD ranges between 20–100mg/L, occasionally reaching up to 34 DO values are usually within 5–7mg/L but can rise to 50mg/L. TDS (maximum 487mg/L), EC (up to 855µS). The heatmap shows the Fig. 4 relationships between the different water quality parameters. A significant positive relation is discovered between: BOD & COD (0.65), COD & OPI (1.00), and TDS & EC (0.70). Most of the remaining correlations are found to be weak, e.g., between pH and COD (–0.13) and DO and EC (–0.07). Additionally, Fig. 5 illustrates the spread of the OPI, exhibiting a positively skewed distribution with most values ranging between 0 and 30. There are many samples with low OPI values and very few such as those with high OPI values that are not common and are outliers.

B. Organic Pollution Index

The OPI deterime the organic contamination in aquatic environments, relying on four critical water quality parameters. It is acting as a combined measure combining the nutrient concentrations and oxygen to give a numeric measure of water quality. The OPI values show high levels of organic contamination making the index useful for comparison and comparison analysis during pollution assessment. The following formula is used to compute the OPI. [19]:

$$\text{OPI} = \frac{\text{COD}}{\text{COD}_s} + \frac{\text{DIN}}{\text{DIN}_s} + \frac{\text{DIP}}{\text{DIP}_s} - \frac{\text{DO}}{\text{DO}_s} \quad (1)$$

In Equation 1:

- **COD** representing Chemical Oxygen Demand, assesses the volume of oxygen necessary to degrade organic compounds within aquatic systems.

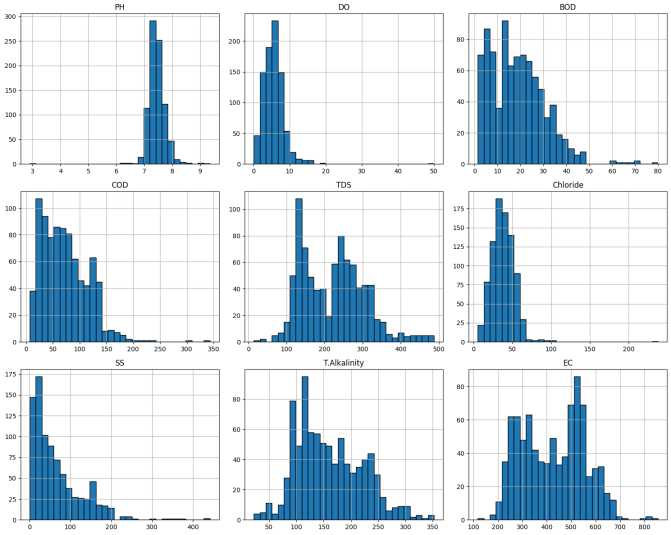


Fig. 3. Distribution of water quality parameters showing skewness, variability, and potential outliers across the sampled data.

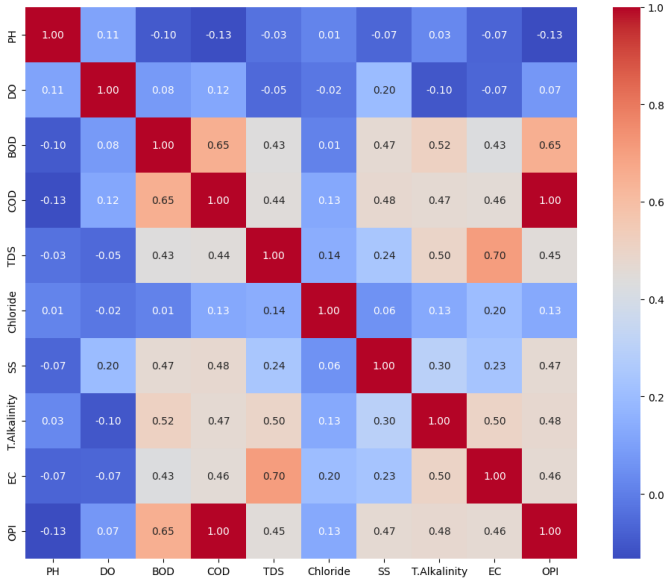


Fig. 4. Correlation Matrix.

- **DIN** is the Dissolved Inorganic Nitrogen concentration, indicating nutrient load from nitrogen compounds.
- **DIP** represents Dissolved Inorganic Phosphorus, another nutrient contributing to eutrophication.
- **DO** stands for Dissolved Oxygen, essential for aquatic life and an inverse indicator of organic pollution.
- The parameters $\mathbf{COD}_s$, $\mathbf{DIN}_s$, $\mathbf{DIP}_s$, and $\mathbf{DO}_s$ denote the *benchmark* concentrations indicative of typical levels found in uncontaminated or pristine water

COD, DIN, DIP and DO are values recorded for the water quality parameters. The clean water typical concentrations for these parameters (COD_s, DIN_s, DIP_s, and DO_s) serve as standard values for these parameters. The calculated OPI value classifies water into six quality categories ranging from "Severely Polluted" to "Pristine," as detailed in Table II.

TABLE II
OPI INDEX AND CORRESPONDING STATUS

OPI Index	Status
5–4	Severely Polluted
4–3	Moderately Contaminated
3–2	Mildly Polluted
2–1	Approaching Contamination
1–0	Beneficial
< 0	Pristine

C. CNN Architecture

The Convolutional Neural Network (CNN) classically has been applied to image processing tasks but recently has posted promising results on tabular and sequential data spaces [20]. A 1D CNN was employed in this work in order to see local patterns as well as interactions among the features of the physico-chemical parameters. All samples were reshaped into the form of 1D sequences with shape (10, 1) so that the convolutional layers can scan over features. CNN architecture consists of two convolutional layers. The initial one applies 64 filters of dimension 3 and ReLU activation to capture efficiently the first-level interactions between neighboring features. The second convolutional layer consists of 32 filters of equal dimension and activation. A flattening process is implemented to prepare data for input into the densely connected neural layers.After processing the gathered information, a dense layer with 64 neurons with the ReLU activation mechanism produces an output layer with a single neuron that predicts the OPI. The model is further compiled using an Adam optimizer since it has an adaptive learning rate and is computationally fast. The CNN model performed training for 100 epochs, with a validation split of 0.2 and a batch size of 32. The capacity of CNN to learn and extract spatially localized patterns ensures that this technique is favorably adapted to model interacting interactions like the inverse correlation between DO and BOD or direct relation of COD on OPI. This modeling process is of special interest in the case of the environmental data where feature interactions are nonlinear and dependent [21].

D. TabNet: Attention-Based Feature Selection

Tab Net is a new deep learning architecture that is expressly proposed for the case of tabular data. It has a decision-step architecture, which uses sparse attention to pick important features at their learning step [22]. This enables the model to specialize dynamically around the critical attributes, without having to do tedious feature re-engineering or manual feature selection. The model applied a decision width (n_d) and attention width (n_a) of 32 units, and five decision steps and a relaxation factor (gamma) equal to 1.5. These hyperparameters govern the amount of reuse and exploration the model does of features. For training the DL model adam optimizer was used where learning rate was set at 0.02 and with the support of the step-wise scheduler to step down the learning rate. Early

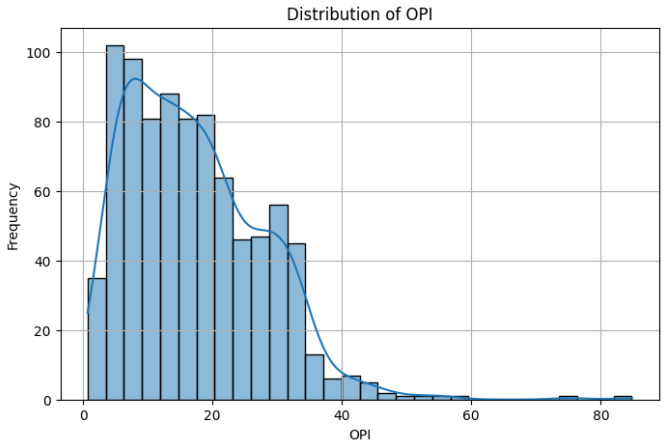


Fig. 5. Right-skewed distribution of OPI values, indicating mostly low to moderate pollution levels.

stopping was used to avoid overfitting, and training ceased when validation loss stopped improving after 20 successive epochs. Batch size was 256, virtual batch size 128 for gradient averaging and training more stable. The interpretability is one of the most important benefits of the TabNet. In the above training the model generates the attention masks that declare which features are the most supportive at any decision step. This heat map transparency enables researchers to see, for example, whether COD, SS or EC values are drawn under conditions of high OPI prediction. This interpretability also makes TabNet an attractive model for studies based on environmental modelling where require explain ability to satisfy regulatory and for public transparency requirements.

E. Hybrid CNN + LSTM Model

Seeing the advantage of integrating both spatial and temporal modeling, an architecture of a hybrid CNN and LSTM model was presented. This architecture is designed to capture local feature interactions as well as temporal pattern from the engineered time features (year, month, day). The model starts with the 1D convolutional layers, 64 of the size 3 filters enclosed in a max-pooling layer in order to dimenions reduction and accentuation common spatial features. Out of the output of the convolutional block goes to an LSTM layer of 64 units. LSTM networks, initially designed for sequence modeling, exhibit the ability to maintain long-term dependencies with gated memory cells and may be used for detecting seasonal patterns in environmental data [23]. To complete the model process are, there are dense layers with 64 neurons and a linear output layer for regression. Notwithstanding the fact that the dataset employed in this research does not have continuous time series data, the hybrid model makes use of temporal characteristics in order to detect recurring pattern of pollution.

F. XGBoost

The ensemble method, Extreme Gradient Boosting (XGBoost), is a very efficient algorithm that also makes use of additive training and regularized learning methodology to avoid over fitting [6]. It builds decision trees one by one, trying to correct the mistakes of the previous one by the next one. Because of its scalability and regularizing abilities, XGBoost has outperformed other ensemble models in structured data competitions. An XGBoost model initiated with 300 estimators, learning rate was set to 0.05 and a maximum of depth was 6. Both subsample and $colsample_bytree$ were initialized to 0.8 ensuring balanced control of overfitting and model diversity. The trained model used the squared error as the objective function. XGBoost's power in dealing with nonlinear relationship and heterogeneous features made it a nice benchmark in this study.

G. Ensemble Model Combining CNN, TabNet, and XGBoost

To improve prediction accuracy and decrease individual model variance, it was decided to implement a soft ensemble method. The last prediction for each testing sample was concluded based on the weighted average output of the CNN, TabNet and XGBoost models. Under this equation ensemble is:

$$\text{OPI}_{\text{ensemble}} = 0.5 \times \text{XGBoost} + 0.3 \times \text{TabNet} + 0.2 \times \text{CNN} \quad (2)$$

Empirically based on verdict of each model on validation set, these weights were chosen. Weights assigned in the ensemble model integrate features learned of structural learning of XGBoost, attention-based feature prioritization of TabNet and spatial interaction learning ability of CNN. This combination is targeted toward stabilizing prediction results and increasing model robustness especially for use in real world lake pollution monitoring systems where noisiness of data and its variability are typical.

V. Results and Discussion

The model's performance was evaluated by using three major measures: MAE, RMSE and Coefficient of determination (R^2 score). Overall, these metrics summarize prediction accuracy, as well as variance sensitivity and model fit. As compared to all other models tested – the CNN model provided the highest accuracy; its MAE was 0.1888, RMSE 0.5930, and R^2 0.9969 as shown in Fig. 6. Table III presents the accuracy of the different DL models evaluated using the metrics MAE, RMSE, MSE, and R^2. These values tell us that the CNN could predict OPI values accurately and with little variance to actual targets. Such result highlights CNN's proficiency to identify and learn valuable spatial relations in the dataset, which is important in modeling environmental variables and where relationships between features such as DO, BOD, and EC have a deep effect on the pollution resulting.

The TabNet model, just below the accuracy level of the CNN, showed great performance with a MAE = 0.6916, RMSE = 1.1451 and an R^2 = 0.9885. What was particularly valuable in this context about TabNet is its interpretability. The attention-based mechanism enabled dynamic prioritization of features and provided the resulting insights about the variables that had the biggest impact at different stages of the process

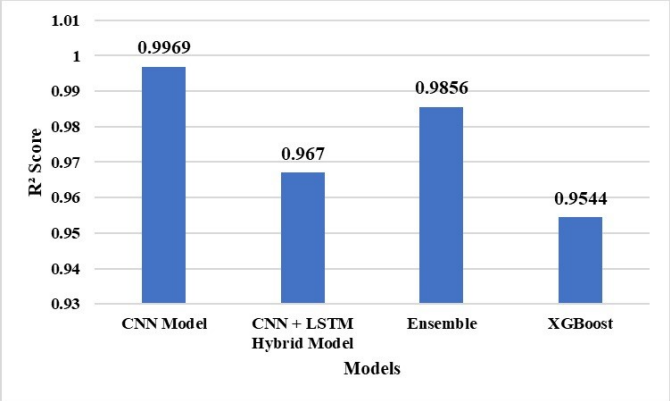


Fig. 6. The bar chart showing R^2 score comparison of Deep Learning models

of prediction. This explainability is essential in environmental science applications where model transparency, as well as the capacity to explain predictions, is as important as accuracy.

TABLE III
PERFORMANCE METRICS OF PREDICTIVE MODELS

Model	MSE	RMSE	MAE	R^2
CNN	0.3516	0.5930	0.1888	0.9969
CNN + LSTM Hybrid	3.7563	1.9381	1.1055	0.9670
Ensemble	1.6370	1.2794	0.5645	0.9856
XGBoost	5.1969	2.2797	0.8868	0.9544

The existing gradient-Boosting framework, the XGBoost Regressor, achieved an MAE = 0.8868, RMSE = 2.2797, and R^2 = 0.9544. While its performance was lower compared to the CNN and TabNet, it served as a reliable classical baseline and exhibited strong generalization capabilities. XGBoost's tree-based structure effectively handled feature interactions and non-linear relationships, although it lacked the fine-grained feature extraction ability seen in deep neural networks. The hybrid CNN + LSTM model produced mixed results. It achieved an MAE = 1.1055, RMSE = 1.9381, and an R^2 = 0.9670. Although this model successfully combined spatial and sequential feature extraction, the absence of truly sequential data (beyond engineered time features) limited the effectiveness of the LSTM layer. This suggests that while the hybrid model has potential, especially in time-series scenarios with more granular or temporally dependent data, it may not be optimal for this specific dataset structure. The ensemble model, which integrated predictions from CNN, TabNet, and XGBoost, demonstrated a balanced performance. It achieved a MAE of 0.5645 and an RMSE of 1.2794, with an estimated R^2 of approximately 0.9800. Although this ensemble did not surpass the CNN in pure accuracy, it offered a stable compromise by reducing individual model biases and providing a smoother error distribution. The ensemble model will be of special use in real-world deployment where robustness under varied conditions is critical.

VI. CONCLUSION

In this study perform an analysis of ensemble and DL algorithms in the prediction of the OPI from physicochemical water quality indicators. The CNN model was the overall best model having an R^2 of 0.9969 and MAE of only 0.1888 which indicates its greater capacity to model complex relationships between environmental variables. TabNet also performed well, with the interpretability benefit, and is a good solution for explainable model use in ecological decision-making. However, the combination of the CNN model with the LSTM model that is theoretically promising could not work in its best assured form, under the absence in the dataset of real temporal dynamics. Although a little bit less accurate, XGBoost provided a very generalisable and interpretable solution which is still a good choice for modelling structured data. The ensemble approach, as CNN, XGBoost, and TabNet were mixed, was a great strong choice after all three individual models applied and gained from their strengths. From a nutshell, the findings of the present study validate the feasibility of AI-driven models for the purpose of monitoring environmental pollution. Through replacing laboratory or manual work with intelligent predictive models, stakeholders can obtain real time scalable low cost solutions for identifying organic pollution in water bodies. Further follow-ups may involve integration of temporal trends, spatial GIS data or remote sensing inputs to achieve even improved predictions and to improve generalizability of the models to various environmental conditions. The limitations of this study include the missing of two important values of water quality such as DIN and DIP, the inaccessible data, the small number of 864 samples and the concentration of 3 lakes in Dhaka city. Future research needs to increase the size of the dataset, add more parameters and use more sophisticated deep learning techniques.

REFERENCES

[1] Mongabay, "Dhaka's ailing sewage system threatens human and environmental health," 2022. [Online]. Available: https://news.mongabay.com/2022/11/dhakas-ailing-sewage-system-threatens-human-and-environmental-health/

[2] M. Ahmed and H. Majeed, "© icas 2022 an assessment of water quality of the north and south part of the shatt al-arab using organic pollution index (opi)," vol. 65, pp. 4939–4945, 01 2022.

[3] T. News, "What is the quality of dhaka's lake waters?" 2022. [Online]. Available: https://www.tbsnews.net/features/panorama/what-quality-dhakas-lake-waters-852411

[4] A. Najah Ahmed, F. Binti Othman, H. Abdulmohsin Afan, R. Khaleel Ibrahim, C. Ming Fai, M. Shabbir Hossain, M. Ehteram, and A. Elshafie, "Machine learning methods for better water quality prediction," *Journal of Hydrology*, vol. 578, p. 124084, Nov. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.jhydrol.2019.124084

[5] S. Talukdar, Shahfahad, S. Ahmed, M. W. Naikoo, A. Rahman, S. Mallik, S. Ningthoujam, S. Bera, and G. Ramana, "Predicting lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms," *Journal of Cleaner Production*, vol. 406, p. 136885, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.jclepro.2023.136885

[6] G. Li, Z. Shu, M. Lin, J. Zhang, X. Yan, and Z. Liu, "Comparison of strategies for multistep-ahead lake water level forecasting using deep learning models," *Journal of Cleaner Production*, vol. 444, p. 141228, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.jclepro.2024.141228

[7] M. Uddin, T. Kormoker, M. A. B. Siddique, M. M. Billah, M. Rokonuzzaman, A. Al Ragib, R. Proshad, M. Y. Hossain, M. K. Haque, K. A. Ibrahim, and A. M. Idris, "An overview on water quality, pollution sources, and associated ecological and human health concerns of the lake water of megacity: a case study on dhaka city lakes in bangladesh," *Urban Water Journal*, vol. 20, no. 3, p. 261–277, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1080/1573062X.2023.2169171

[8] M. Hassan, "Sewage waste water characteristics and its management in urban areas- a case study at pagla sewage treatment plant, dhaka," *Urban and Regional Planning*, vol. 2, no. 3, p. 13, 2017. [Online]. Available: http://dx.doi.org/10.11648/j.urp.20170203.11

[9] R. Mezbour, R. Asma, M.-C. Maazi, and M. Houhamdi, "Evaluation of organic pollution index and the bacteriological quality of the water of the lake of birds (eltarf east-algerian)," *Journal of Materials and Environmental Science*, vol. 9, pp. 971–979, 01 2018.

[10] A. Al-Asadi and M. Al-Hejuje, "Application of organic pollution index (opi) to assess the water quality of al-chibayish marsh, southern iraq," vol. 14, pp. 57–64, 07 2019.

[11] R. M. Frincu, "Artificial intelligence in water quality monitoring: A review of water quality assessment applications," *Water Quality Research Journal*, vol. 60, no. 1, p. 164–176, Nov. 2024. [Online]. Available: http://dx.doi.org/10.2166/wqrj.2024.049

[12] W. Luo, L. Huang, J. Shu, H. Feng, W. Guo, K. Xia, K. Fang, and W. Wang, "Predicting water quality in municipal water management systems using a hybrid deep learning model," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108420, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2024.108420

[13] S.-S. Baek, J. Pyo, and J. A. Chun, "Prediction of water level and water quality using a cnn-lstm combined deep learning approach," *Water*, vol. 12, no. 12, p. 3399, Dec. 2020. [Online]. Available: http://dx.doi.org/10.3390/w12123399

[14] Z. Mi, Q. Li, Y. Sha, and Z. Wu, "Cnn-lstm-attention water quality prediction hybrid model," in *Second International Conference on Sustainable Technology and Management (ICSTM 2023)*, X. Qu, Ed. SPIE, Oct. 2023, p. 5. [Online]. Available: http://dx.doi.org/10.1117/12.2684594

[15] G. Wu and C. Zhang, "Analysis of water quality prediction in the yangtze river delta under the river chief system," *Sustainability*, vol. 16, no. 13, p. 5578, Jun. 2024. [Online]. Available: http://dx.doi.org/10.3390/su16135578

[16] Z. Wei, L. Wei, H. Yang, Z. Wang, Z. Xiao, Z. Li, Y. Yang, and G. Xu, "Water quality grade identification for lakes in middle reaches of yangtze river using landsat-8 data with deep neural networks (dnn) model," *Remote Sensing*, vol. 14, no. 24, p. 6238, Dec. 2022. [Online]. Available: http://dx.doi.org/10.3390/rs14246238

[17] M. Uddin, T. Kormoker, M. A. B. Siddique, M. M. Billah, M. Rokonuzzaman, A. Al Ragib, R. Proshad, M. Y. Hossain, M. K. Haque, K. A. Ibrahim, and A. M. Idris, "An overview on water quality, pollution sources, and associated ecological and human health concerns of the lake water of megacity: a case study on dhaka city lakes in bangladesh," *Urban Water Journal*, vol. 20, no. 3, p. 261–277, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1080/1573062X.2023.2169171

[18] J. Brownlee. (2022) Data preparation for machine learning. [Online]. Available: https://machinelearningmastery.com/data-preparation-for-machine-learning/

[19] C.-W. Chen, Y.-R. Ju, C.-F. Chen, and C.-D. Dong, "Evaluation of organic pollution and eutrophication status of kaohsiung harbor, taiwan," *International Biodeterioration amp; Biodegradation*, vol. 113, p. 318–324, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.ibiod.2016.03.024

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436–444, May 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14539

[21] U. Ewuzie, O. P. Bolade, and A. O. Egbedina, *Application of deep learning and machine learning methods in water quality modeling and prediction: a review*. Elsevier, 2022, p. 185–218. [Online]. Available: http://dx.doi.org/10.1016/B978-0-323-85597-6.00020-3

[22] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," 2019. [Online]. Available: https://arxiv.org/abs/1908.07442

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735